

# Compatibility, Cliques and Clonal Frames

Barbara Holland  
University of Tasmania



# Unravelling the processes of bacterial evolution

- Processes

- Mutation



- Homologous recombination



- HGT

- Data is available at multiple levels of resolution

- Gene presence / absence

- Allele profile

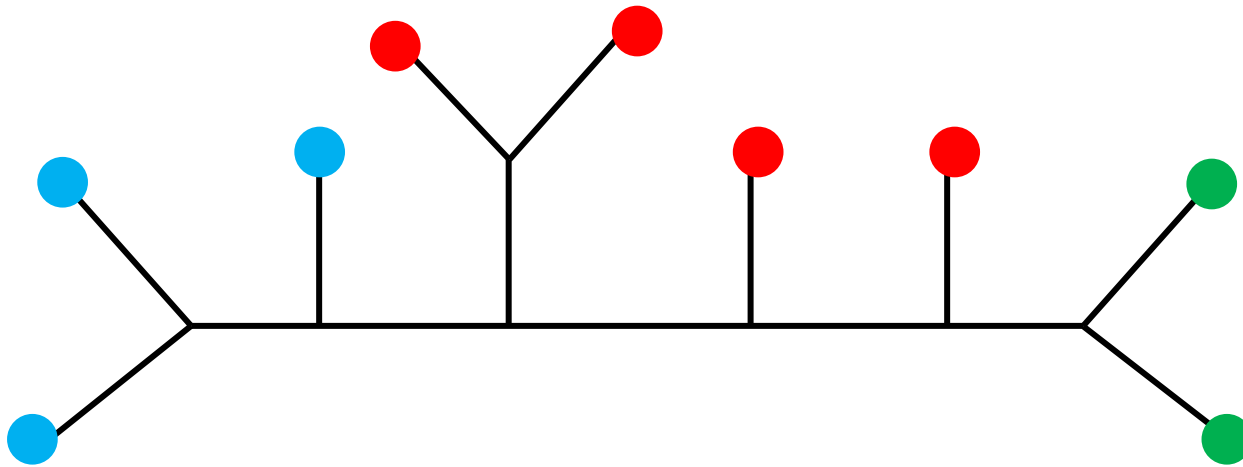


- Sequence data



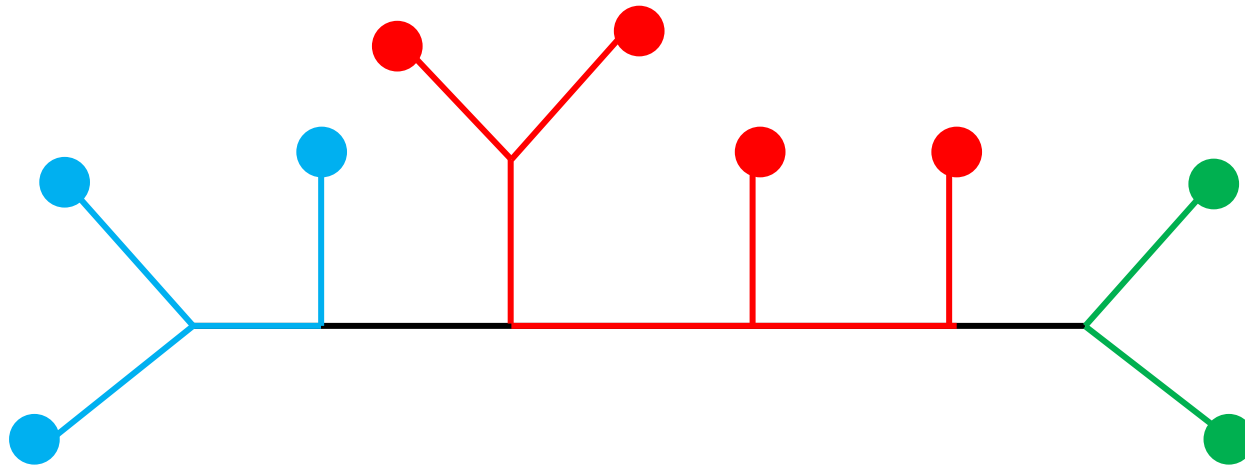
# Compatibility

Given a character  $C$  and a tree  $T$  we can ask if the character is compatible with the tree.



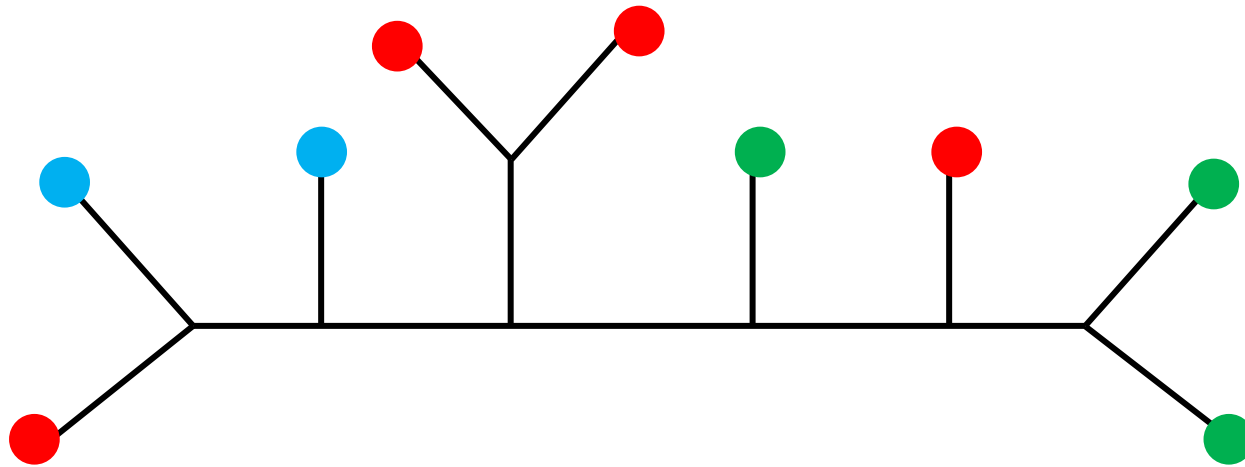
# Compatibility

Given a character  $C$  and a tree  $T$  we can ask if the character is compatible with the tree.

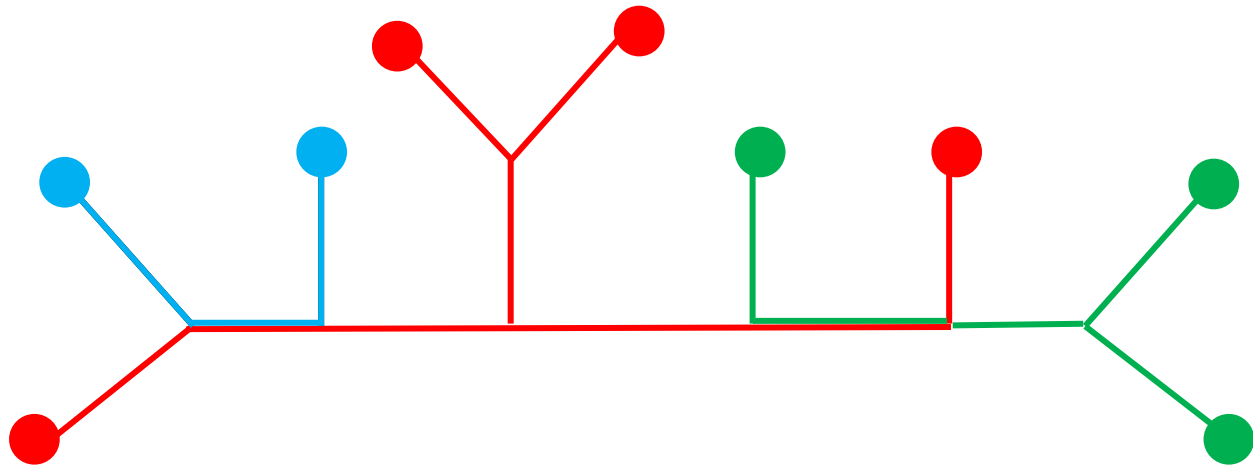


A compatible character

# Incompatibility



# Incompatibility



An incompatible character

# Compatible cliques of characters

- Characters are said to be compatible with each other if there exists a tree which they are all compatible with.

# Allele profile data

- Multi-level data
  - Strain type
  - Allele profile
  - Sequence

e.g. MLST data

locus	L1	L2	L3	L4	L5	L6	L7
ST1	1	1	1	1	1	1	1
ST2	1	1	2	1	1	1	1
....							



## L3

1	CCCTTGTTTAGTCCAAATTCACACCAATTTCA
2	CCCTT <b>A</b> TTTAGTCCAAATTCACACCAATTTCA
...	...

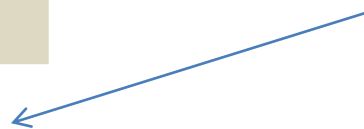


# Allele profile data

- Multi-level data
  - Strain type
  - Allele profile
  - Sequence

e.g. MLST data

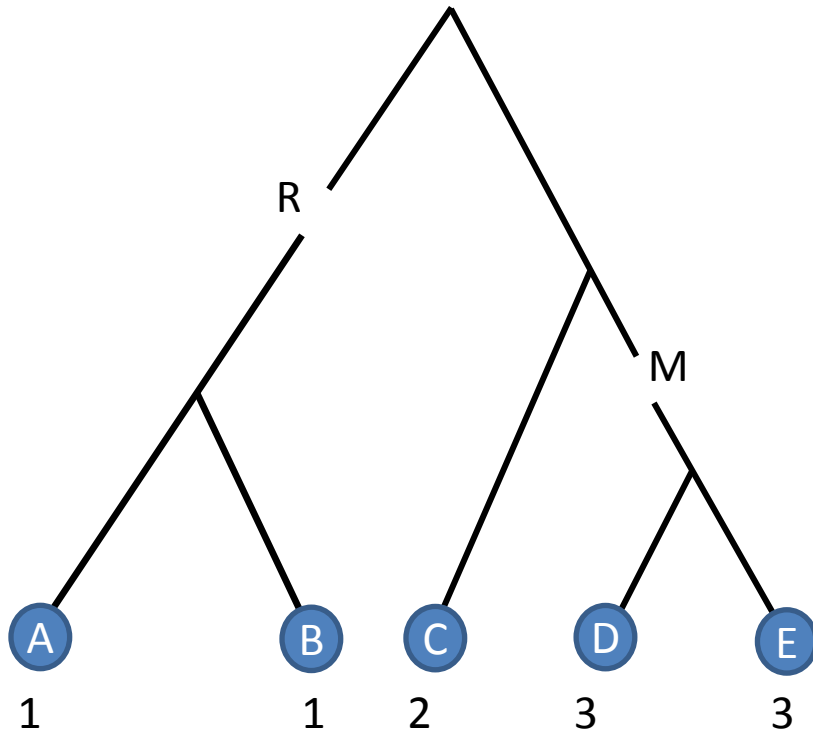
locus	L1	L2	L3	L4	L5	L6	L7
ST1	1	1	1	1	1	1	1
ST2	1	1	2	1	1	1	1
....							



## L3

1	CCCTTGTTTAGTCCAAATTCACACCAATTTCA
2	CCCTT <b>A</b> T <b>C</b> T <b>G</b> G <b>C</b> TCAAATTCACACCAATTTCA
...	...

### Clonal Frame



Evolution of a single **locus** along a clonal frame by mutation (M) and recombination (R) events. A locus is a contiguous stretch of DNA – it will be represented by one column in an allele profile.

### Allele types

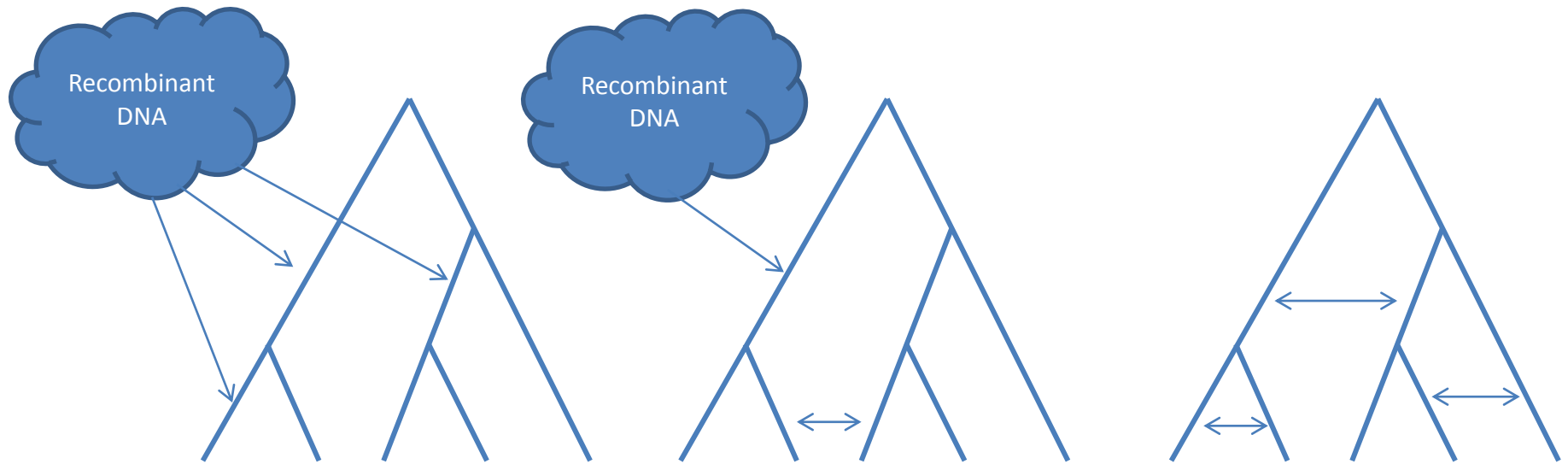
1      ACCG**ATAT**AGGAT**TCGTTCGT**CA  
 2      ACCGTTGCAGGACTGCTAGCCA  
 3      ACCGTTGCAGG**T**CTGCTAGCCA

Allele type 2 and 3 differ from each other in a single position due to a mutation event. Allele type 1 and 2 differ from each other in many positions due to a recombination event. This locus makes up a single column (bold) of the allele profile below.

### Allele Profile

A      1**1**111...  
 B      1**1**212...  
 C      1**2**113...  
 D      2**3**114...  
 E      2**3**114...

# A range of recombination models



(A) ClonalFrame model:  
Recombination always  
introduces novel genetic  
material.

(B) Intermediate model



(C) ClonalOrigin model:  
Recombination always  
occurs within a closed  
population.

Open system



Closed system

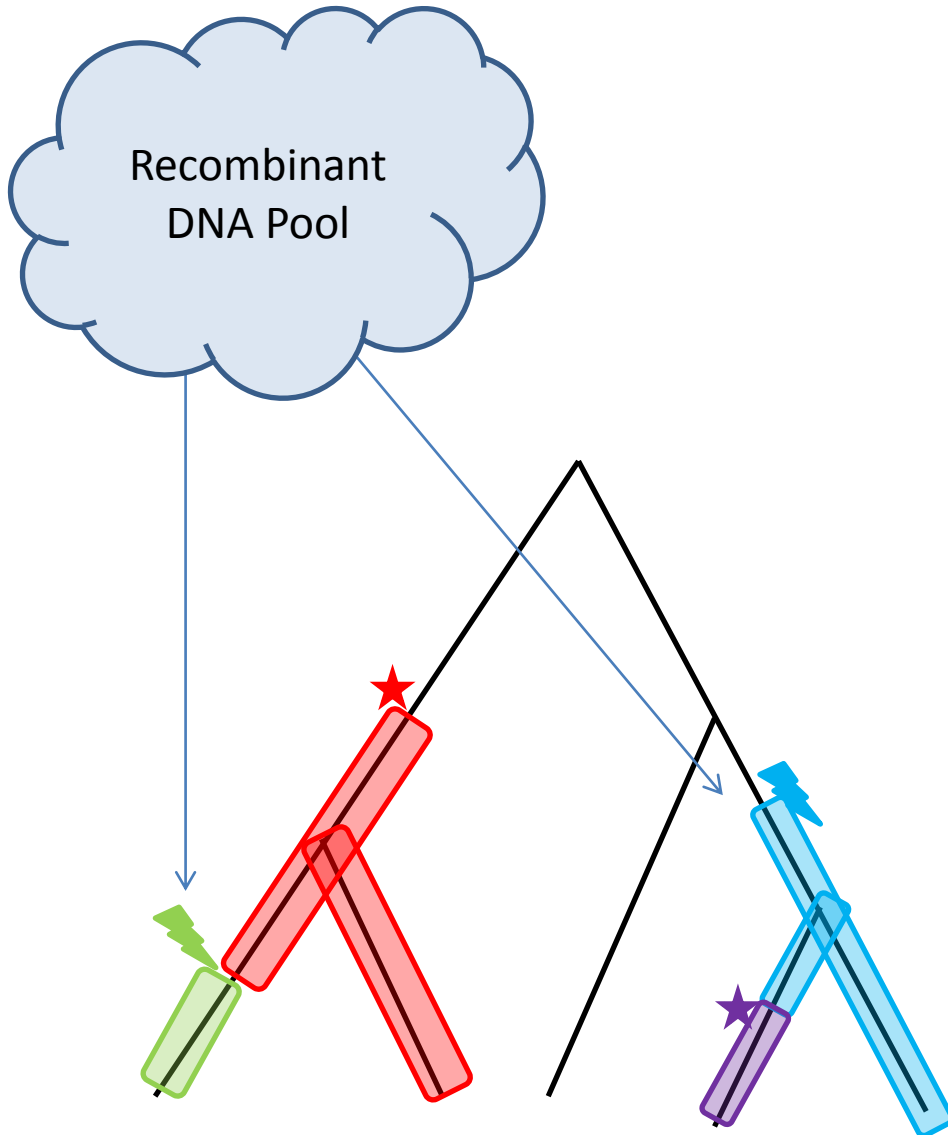
## Clonal Frame model – Infinite Alleles Model?

A particular locus can undergo two types of events  
mutation   
recombination 

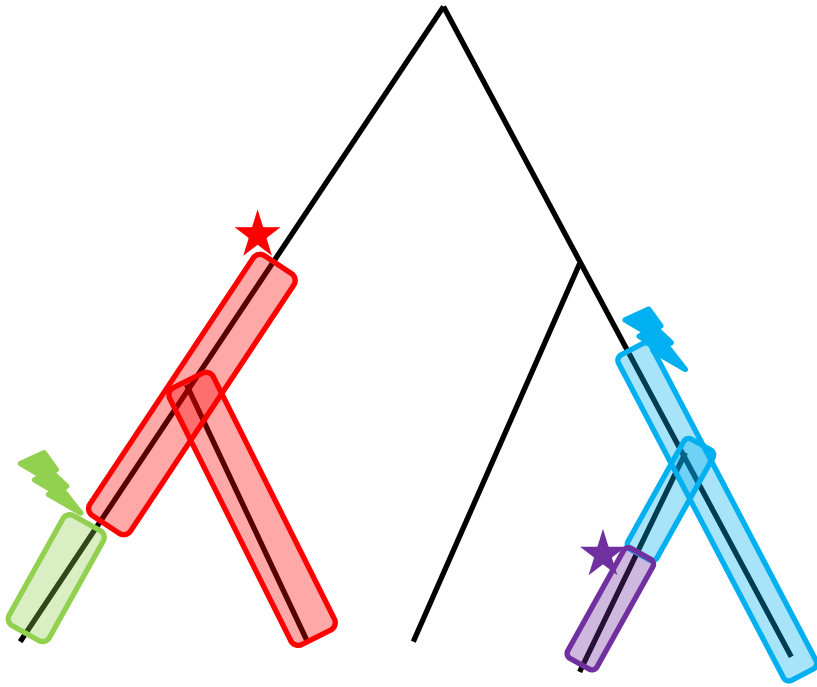
Parallel mutation should be infrequent as it requires

- 1) that the next mutation in the sequence for that locus occurs at the same site, i.e. without any other mutations occurring in the meantime  $p \propto \frac{1}{L}$
- 2) And it further requires that the mutation is back to the initial state

Parallel recombination might be more likely, especially in a closed system. In an open system – as per the ClonalFrame model – parallel recombination should be even less likely than parallel mutation.

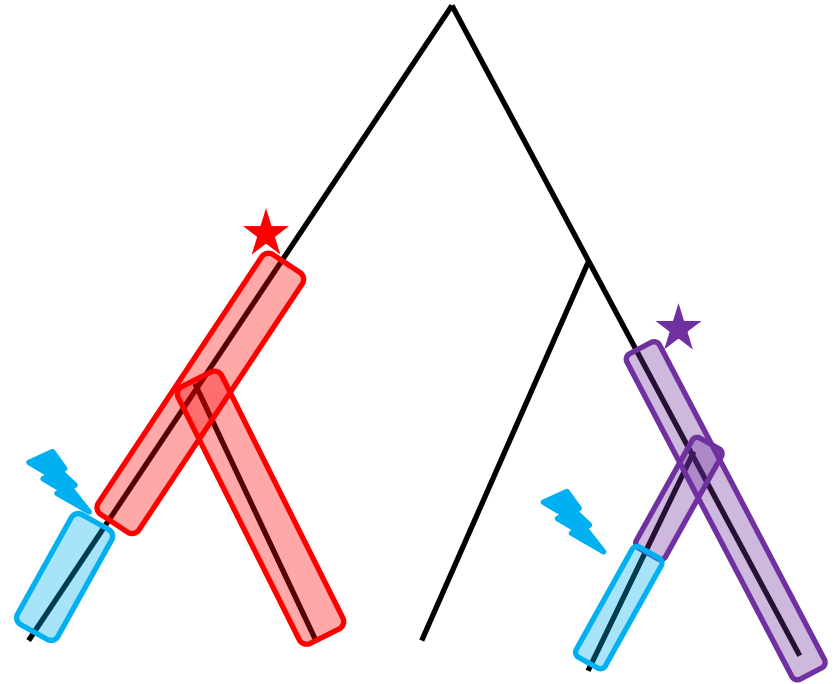


A compatible character



Loci that haven't undergone parallel recombination will produce a character (i.e. a column in the allele profile) that is compatible with the clonal frame.

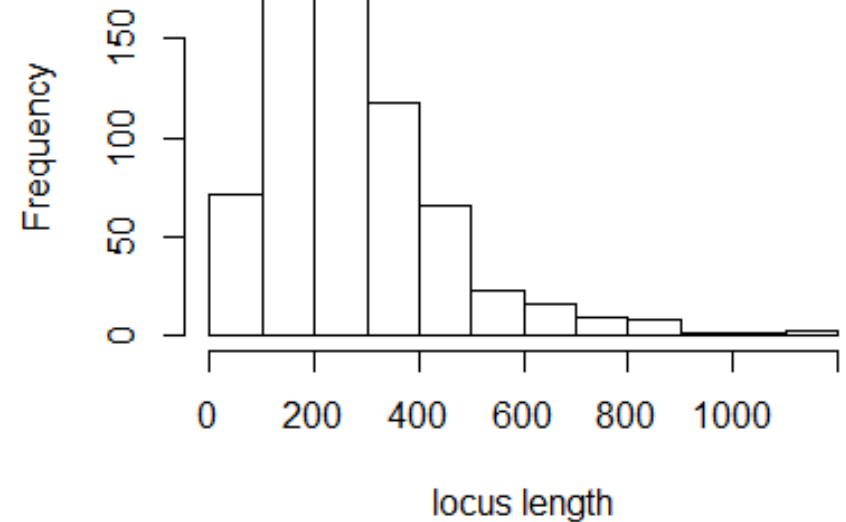
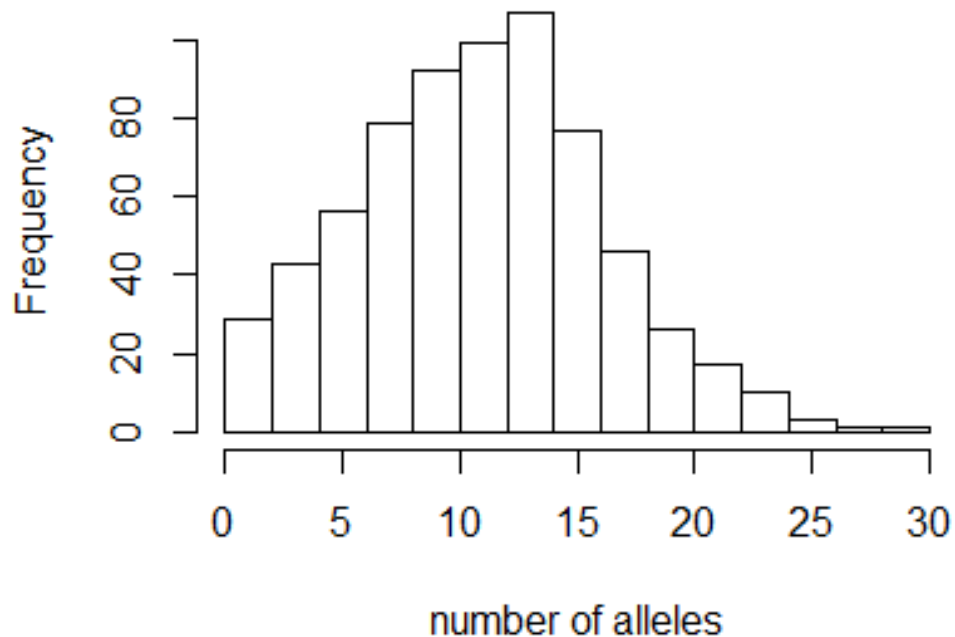
An incompatible character



Blocks that have undergone parallel recombination (or parallel mutation) may produce characters that are not compatible with the clonal frame.

# The *Campylobacter jejuni* data

- 46 *C. jejuni* genomes
- 686 genes in common across all 46 genomes



# Initial analysis

- 686 characters
- 9 constant, 2 parsimony uninformative
- Theoretical best parsimony score 7083

$$\sum_{l=1}^{686} (r_l - 1)$$

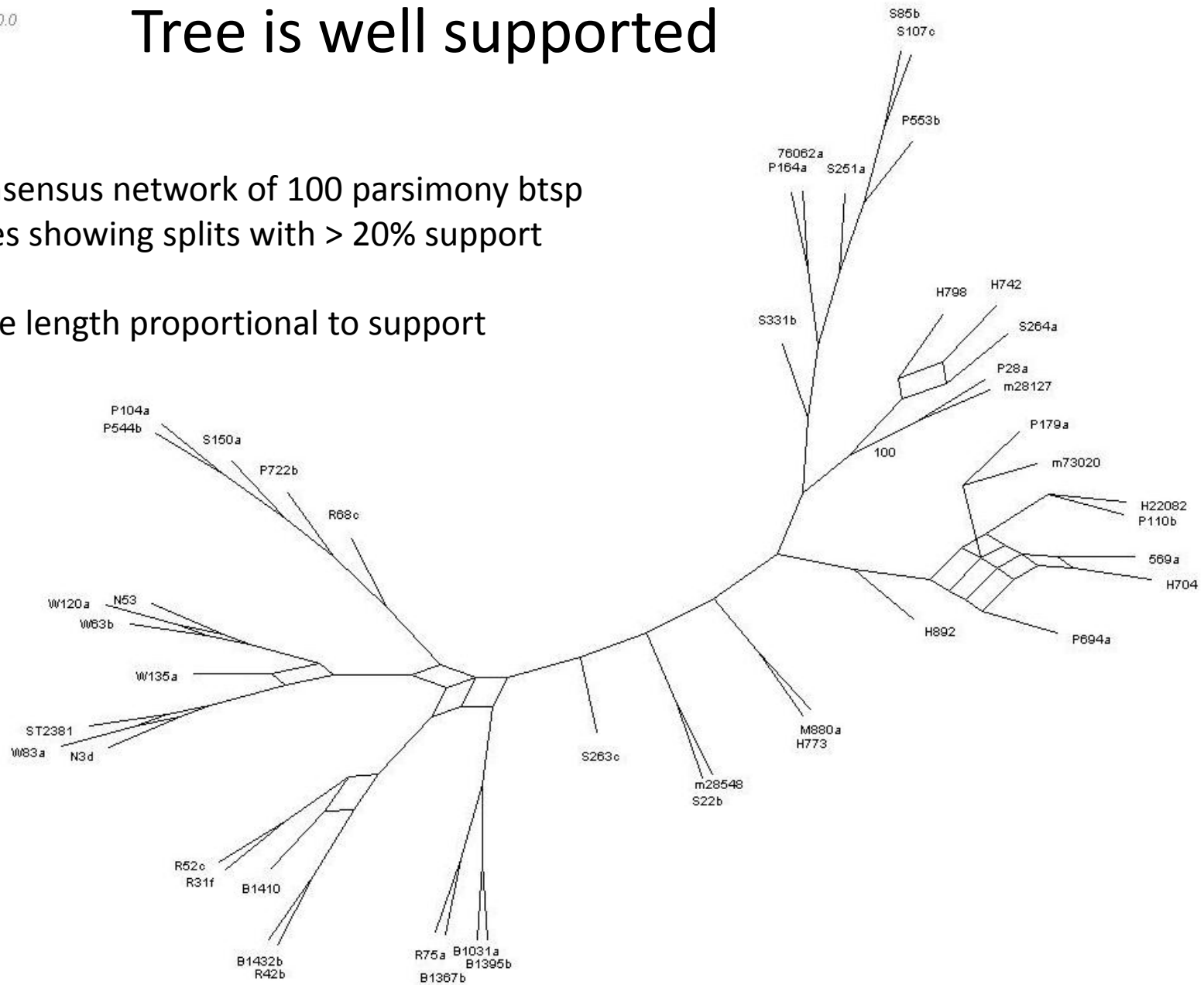
Where  $r_l$  is the number of alleles at locus  $l$

- Parsimony finds 3 equally parsimonious trees with score 8274
- Consistency index 0.856

# Tree is well supported

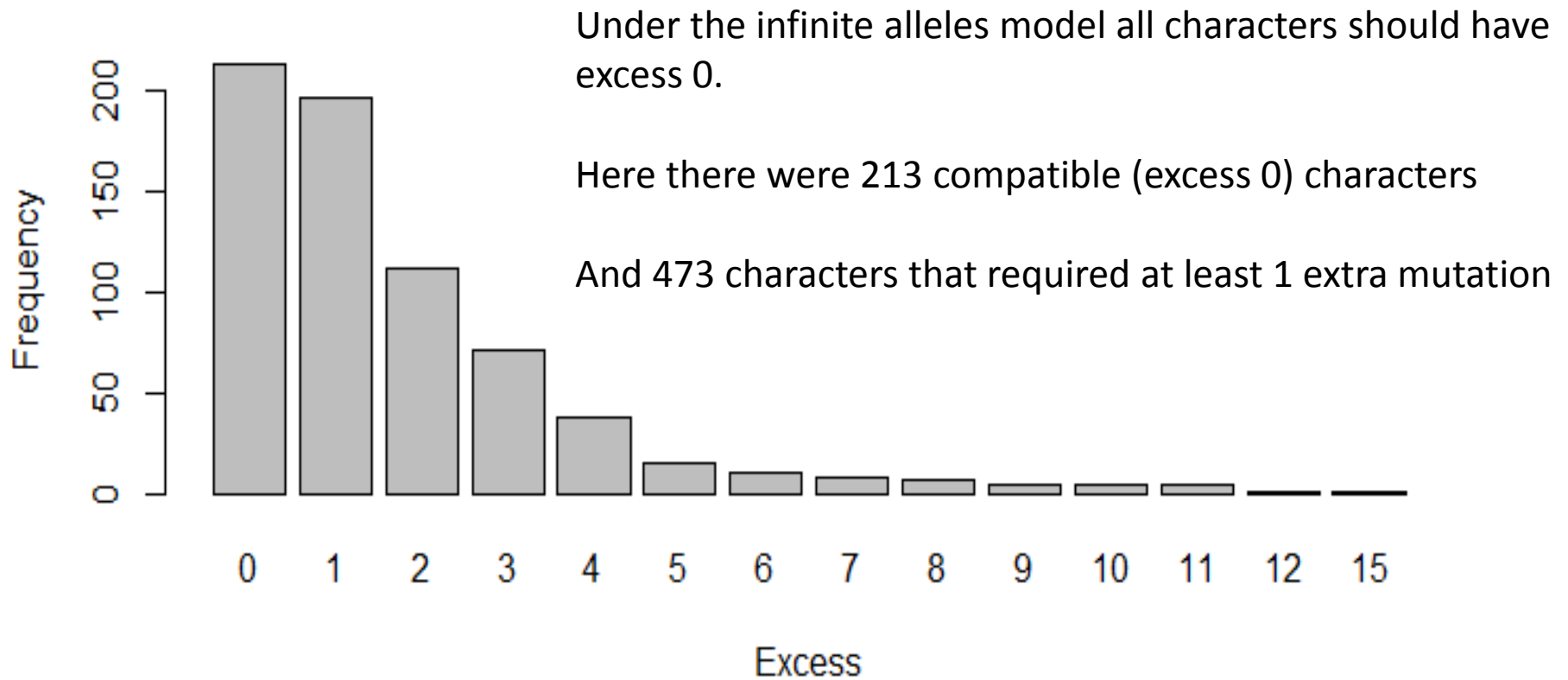
Consensus network of 100 parsimony btsp trees showing splits with > 20% support

Edge length proportional to support





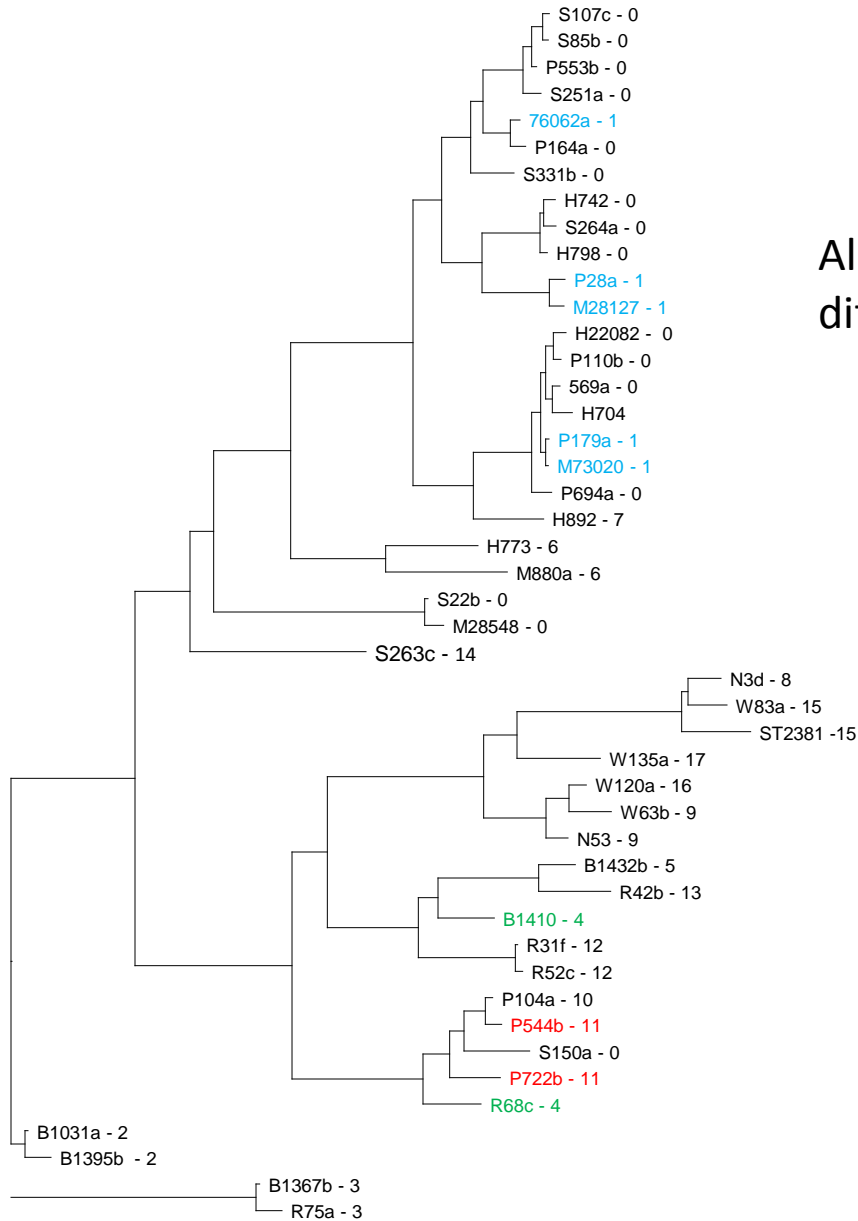
# *Are some genes more prone to parallel events?*



# Ancestral state reconstruction

- Find the clonal frame using maximum parsimony
- Use parsimony version of ASR work out all the transitions from one allele to another – look at the distribution of differences between pairs of alleles.
- Compare the distribution of allele differences of compatible characters to that of incompatible characters

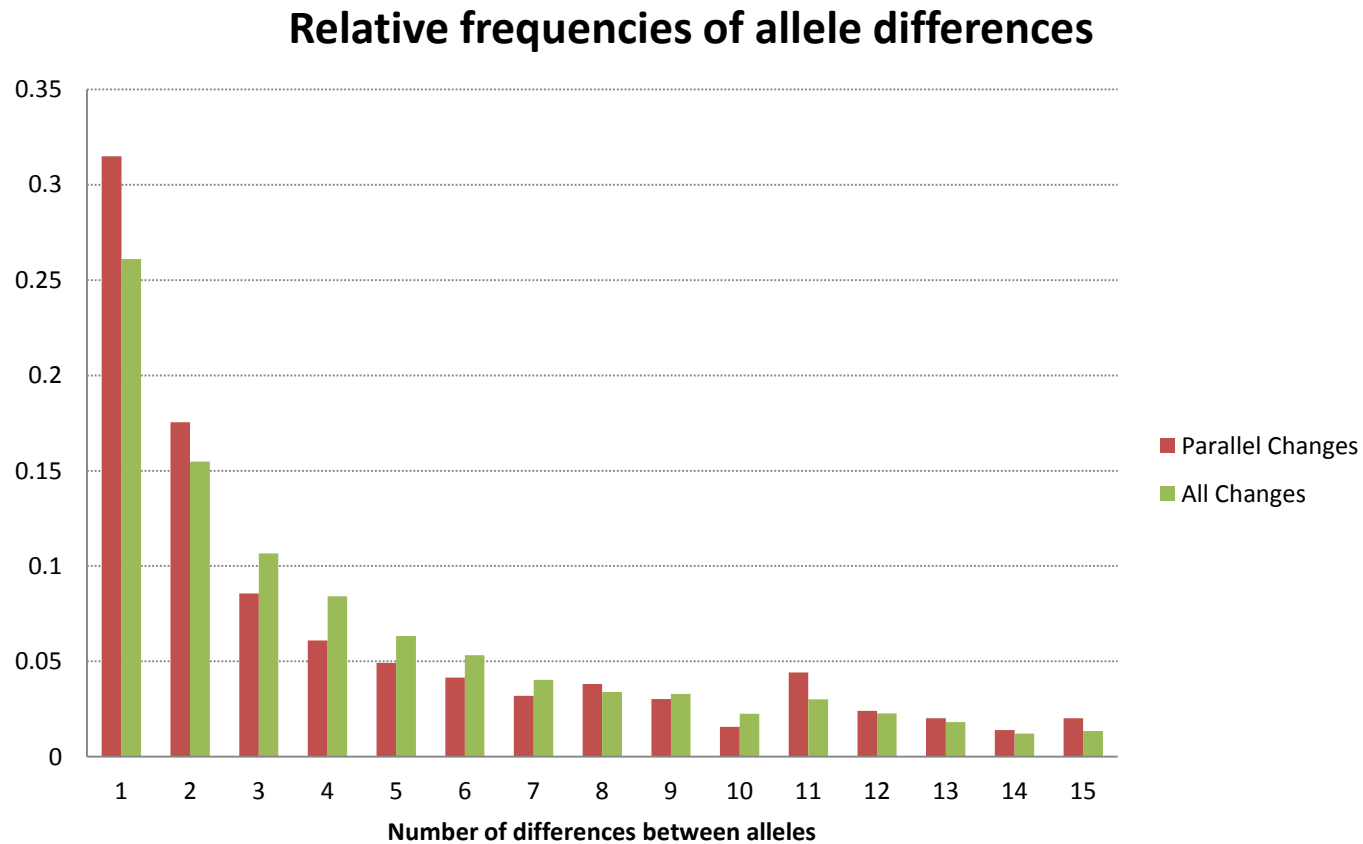
# Clear cases of parallel recombination



Allele 0 and 1  
differ at 20 sites

100185noOut.fa  
18 alleles  
Excess of 3

# *Are parallel events more often mutation or recombination?*



*Are some edges more prone to recombination events?*

- See scribbles

Tree rooted fairly arbitrarily (mid-point attempt by eye)  
 Edges lengths not to scale in this picture

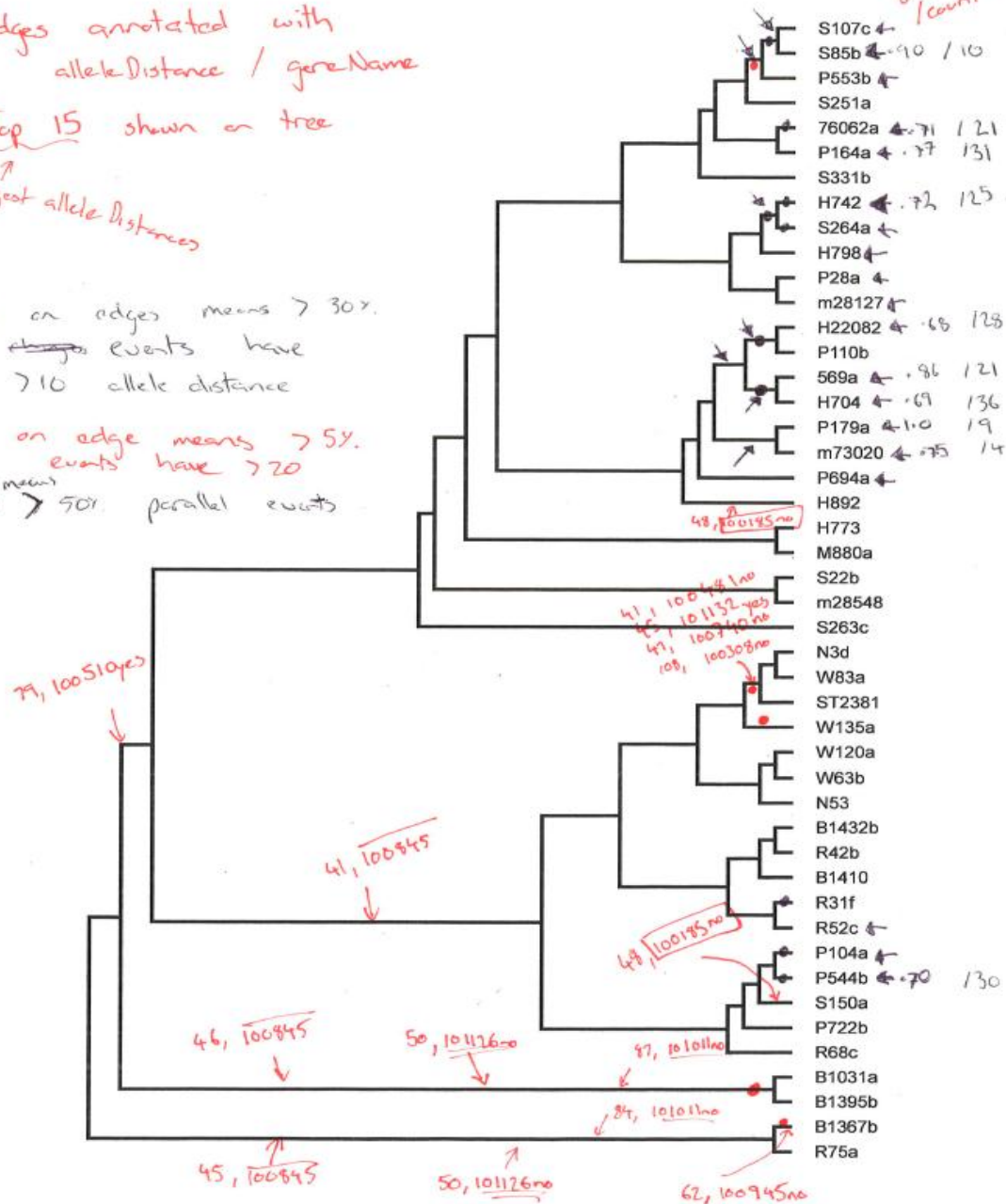
prop. parallel events (top 10) / count

Edges annotated with allele distance / geneName

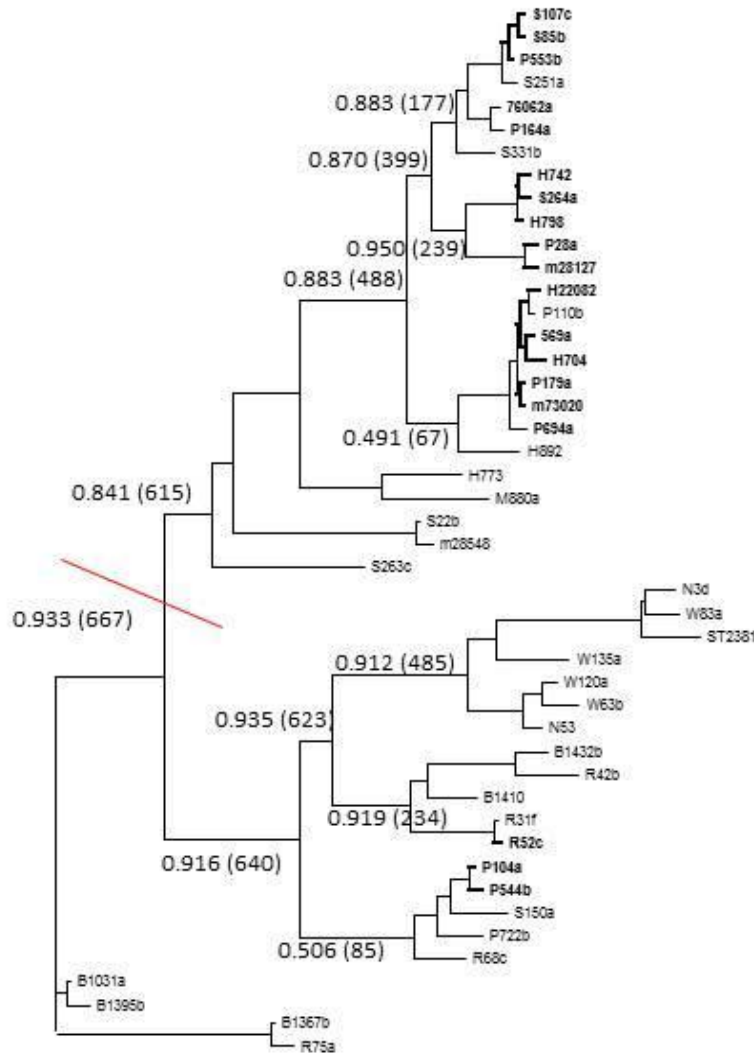
Top 15 shown on tree

↑ largest allele distances

- on edges means > 30% events have > 10 allele distance
- on edge means > 5% events have > 20
- means > 50% parallel events



# Are some edges or clades more prone to parallel events?



Bold edges / labels indicate that more than 50% of events allocated to that edge are parallel events.

Retention index by clade (#parsimony inf. characters)

# Conclusions

- Overall AP data is very consistent, i.e. highly compatible, consistency index  $> 0.85$
- Clonal Frame wastes a lot of computational effort on finding the clonal frame but its model predicts (close to) perfect phylogenies.
- Hard to tell if parallel mutation is more common than parallel recombination as recombination might occur frequently between alleles that aren't very different.
- Seems like different processes predominate in different parts of the tree. Sampling artefact? Testable?



# Acknowledgements

- Nigel French, Patrick Biggs, Shoukai Yu
- Marsden Fund grant to NF